Designing and detecting lies by reasoning about other agents

Lauren A. Oey*, Adena Schachner, & Edward Vul

Department of Psychology

University of California, San Diego

9500 Gilman Drive

La Jolla, CA 92093-0109

Tel: (858) 534-2947; Fax: (858) 534-7190

Author Note

*Previous versions of this work have been presented at the Cognitive Science Society and Society for Philosophy and Psychology conferences (Oey, Schachner, & Vul, 2019a, 2019b; Oey & Vul, 2021). The data and analyses are publicly available on `https://github.com/la-oey/RationalLying` and `https://osf.io/x6rhs/`(Oey et al., 2022, May 25). Correspondence should be addressed to Lauren Oey. E-mail: loey@ucsd.edu

Draft version 2.1 5/26/22.
This paper has not been peer reviewed.

Word Count: 10,033

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1650112 to LAO.

Abstract

How do people detect lies from the content of messages, and design lies that go undetected? Lying requires strategic reasoning about how others think and respond. We propose a unified framework underlying lie design and detection, formalized as recursive social reasoning. Senders design lies by inferring the likelihood the receiver detects potential lies; receivers detect lies by inferring if and how the sender would lie. Under this framework, we can predict the rate and content of lies people produce, and which lies are detected. In Experiment 1, we show that people calibrate the extremeness of their lies and what lies they detect to beliefs about goals and the statistics of the world. In Experiment 2, we present stronger diagnostic evidence for the function of social reasoning in lying: people cater their lies to their audience, even when their audience's beliefs differ from their own. We conclude that recursive and rational social reasoning is a key cognitive process underlying how people communicate in adversarial settings.

Keywords: deception; rational inference; social cognition

Designing and detecting lies by reasoning about other agents

Human communication, though generally honest, is riddled with deception. Most theories of effective communication are predicated on the assumption that interlocutors act cooperatively (Grice, 1975; Grice, 1989). However, in police interviews (Mann et al., 2004), online dating (Hancock & Toma, 2009; Toma et al., 2008), scientific reporting (Fanelli, 2009; John et al., 2012), and news stories (Allcot & Gentzkow, 2017; Lazer et al., 2018), people may choose to present false information. We focus on *lying*, defined here as a sender producing a knowingly false message intended to deceive a receiver. This definition of lying encompasses both verbal and non-verbal communication (Zuckerman et al., 1981) and emphasizes the salient communicative role of the receiver in lying—receivers can believe a lie, or not.

We propose that lying and lie detection arise from interactive, adversarial reasoning where interlocutors must consider how the other will act. In such dyadic communication, receivers are not merely passive audiences—rather, they may punish dishonesty (Ohtsubo et al., 2010; Tyler et al., 2006). Therefore, senders, in deciding which lies to tell, are motivated to avoid being caught by the receiver. Similarly, false accusations are detrimental, and receivers want to avoid them when deciding which messages to call out as lies.

The central idea behind this framework is that the interaction between the competing goals of sender and receiver is critical for deception. While some prior theories highlight how dynamic interaction plays out over the course of back-and-forth conversational sparring (Buller & Burgoon, 1996), our framework highlights the role of anticipated interactivity in human lying cognition, even before a lie is uttered, and places theory of mind (ToM), or the ability to reason about others' mental states and goals (Premack & Woodruff, 1978), at the core of deception. The decision to lie (vs. tell the truth) is known to require some basic ToM understanding, for instance, to acknowledge that receivers may not have access to ground truth and could thus, in principle, be deceived (e.g. Ding et al., 2015). However, the extent to which ToM reasoning drives *how* people actually lie and detect lies has been relatively unexplored.

Theory of mind reasoning is computationally expensive, so people may prefer to rely on other cognitive mechanisms, even when at risk for detection. First, lying is cognitively

demanding (Vrij et al., 2006) and incurs longer response times than telling the truth, even without the risk of getting caught (Capraro et al., 2019; Suchotzki et al., 2017, but see Shalvi et al., 2012). If ToM reasoning itself is a non-automatic, effortful process (Apperly et al., 2006; Lin et al., 2010; Phillips et al., 2015), then applying a complex ToM process would further increase the cognitive demand required of lying. Second, people have been shown to be practically at chance when detecting lies (e.g. Bond & DePaulo, 2006). Under a blanket assumption that detectors are simply guessing, liars need not attribute sophisticated reasoning to lie detectors to succeed at duping them. Third, given the scarcity of distinguishing information about others' idiosyncratic beliefs, a heuristic that relies only on the speakers' own beliefs to choose a lie may well be globally optimal.

Lying and Lie Detection in Isolation and in Dyads

A majority of prior work on lying has omitted key elements of dyadic, adversarial communication that might require theory of mind reasoning, instead focusing on lying (e.g. Gerlach et al., 2019; Mazar et al., 2008) and lie detection (e.g. Bond & DePaulo, 2006; Vrij et al., 2019) in isolation. As a consequence, this prior work cannot speak to whether liars and detectors adapt their strategies in light of considerations of what the other will do. Specifically, research on lie detection has directed its focus on surface features of lies, not the informational content, and has thus paid less attention to the importance of designing lies to be believable. Lie production research, in turn, has primarily used scenarios where liars face no risk of being caught; and thus has also overlooked that real-world lies need to be designed to minimize this risk. In short, by studying lying and lie detection in isolation, prior research has not explored how the two jointly constrain the design of lies in a single communicative act.

Studies of *lie detection* have concentrated on detecting lies from superficial cues, rather than the content of the lie. Classic research on lie detection asked whether lies can be identified from content-independent perceptual cues given off by the speaker, like facial expressions (Bruer et al., 2019; DePaulo et al., 2003; Ekman & Friesen, 1969; Ekman et al., 1988) and verbal pauses (Granhag & Strömwall, 2002; Vrij, 2008). Other research has

prominently been concerned with simple perceptual cues of the message, like whether the statement is repeated (Brashier & Marsh, 2020; Dechêne et al., 2010) or its readability (e.g. statements in **high contrast** are judged as truer than those in low contrast; Reber and Schwarz, 1999; Withall and Sagi, 2021). These extensive bodies of literature have established that content-free perceptual cues to deception are weak and unreliable, despite people's tendency to over-rely on them and their persisting meta-cognitive theories about the diagnosticity of these cues (Vrij et al., 2019). Beyond perceptual cues, other work has focused on lie detection from social and contextual cues, including another person's incentive to lie (Bond et al., 2013; Kraut, 1978), or the (low) base-rate of lying, used as a proxy for making truth judgments (Levine, 2014; Street, 2015). This prior research on lie detection has not examined the relationship between the content of the lie, the receiver's prior knowledge of the world, and their beliefs about the sender's cognitive processes.

Meanwhile, studies of *lying* have examined behavior in scenarios with no risk of being caught, including how (un)willing people are to lie or cheat (Abeler et al., 2019; Gerlach et al., 2019; Mazar et al., 2008), how liars respond to incentives (Gneezy et al., 2013; Mazar et al., 2008), and what lies people produce (Fischbacher & Föllmi-Heusi, 2013; Hilbig & Hessler, 2013; Shalvi et al., 2011). This implicit idea that dyadic interaction is not needed to understand lying behavior is made explicit in the self-concept maintenance account, which proposes that people's lies are constrained by their own beliefs, e.g. about their own honesty (Gino et al., 2009) and moral virtue (Mazar et al., 2008). This account was designed specifically to explain why, even in situations without the risk of detection, people seems to avoid producing large lies. This work posits that aversion to lying, and the selection of lies, is guided by heuristics internal to the speaker.

In contrast to this idea, recent work using a dyadic approach targets how people may consider the listener when lying. These accounts adopt game-theoretic approaches to model strategic behavior in adversarial situations (Becker, 1968). Dyadic frameworks have succeeded at explaining systematic human preferences for general deceptive strategies (e.g. Montague et al., 2011). For example, research in this vein has shown that senders generally prefer to mislead over outright lying, but when receivers are suspicious, senders elect to be

uninformative (Franke et al., 2020; Ransom et al., 2019). Game-theoretic approaches also explain indirect speech for soliciting bribes in circumstances when the speaker is uncertain about the audience's disposition (J. J. Lee & Pinker, 2010). And prominently, lying is more prevalent when it may benefit the listener (as in "white lies": Erat and Gneezy, 2012; Gneezy, 2005), and is less prevalent when others are able to verify the ground truth (Gneezy et al., 2018). This work suggests that reasoning about the interlocutor as having a goal or belief at all—i.e. some kind of theory of mind—may play a key role in lying.

While there is a growing body of literature on dyadic frameworks for understanding deception, many of these studies are designed to understand strategies *other than lying*. Several studies have focused on misleading information (i.e. strategically uninformative content) by considering settings where senders are explicitly prevented from lying (Ransom et al., 2019), or are provided no incentives to lie rather than just mislead (Montague et al., 2011; Rogers et al., 2017). In these cases, lies are unnecessary, and so there is no motive to design them well, or use them at all. On the other hand, most studies of lying have used settings where speakers are not punished for being caught in a lie (e.g. Gneezy, 2005, but see Gneezy et al., 2018), thus again removing incentives to lie strategically. The net effect is that existing research on lying has not explored scenarios where speakers are motivated to design lies that are both advantageous to the sender and plausible to the receiver.

Here we propose, and test, an account of lying as a fundamentally dyadic adversarial reasoning problem. We posit that people detect and generate lies in adversarial settings, by selecting counter-strategies tailored to the behavior of the opponent that they predict, all under the assumption that the opponent is a rational, thinking agent with particular goals and knowledge of the world. We formalize this account in recursively coupled, adversarial theory of mind models of the liar and lie detector. We introduce a novel dyadic lying game, allowing us to measure and parametrically manipulate lying and lie detection behavior in an adversarial context. This experimental context allows us to test whether people lie and detect lies by reasoning about other agents. In Experiment 1, we use this paradigm to test whether senders consider receivers' beliefs and adjust the plausibility of their lies to the statistics of the world; while receivers reason about the senders' goals, and thus rationally adjust which claims they

call out as lies. In Experiment 2, we further test whether senders adapt specifically to the statistics of the world they think that receivers believe to be true, even when they know these beliefs to be false; thus testing the central role of theory of mind representations in the strategic design of lies. Altogether, we find that human behavior exhibits the key qualitative patterns of adversarial theory of mind reasoning predicted by our formal model.

Formalizing Dyadic Reasoning in Lying and Lie Detection

As a first step, we introduce a formal model of dyadic reasoning in lying and lie detection. Formal models allow us to explicitly define our cognitive assumptions and generate behavioral predictions, which we can empirically test. Most importantly, to test whether dyadic reasoning is driving the behavioral predictions, we can compare the predictions of the dyadic reasoning model to those of alternative models that drop the critical theory of mind reasoning assumptions.

To formalize the interactive, adversarial reasoning inherent in lying and lie detection, we develop an ideal observer model inspired by recursive probabilistic inference models of human social cognition and cooperative communication (Frank & Goodman, 2012; Kao et al., 2014; Shafto et al., 2014). In this account, a *sender S* chooses what to say in light of how they believe a *receiver R* will respond, and the receiver decides whether the utterance is a lie based on what they believe a sender would say in different world states. We formalize the lying interaction as follows: the sender observes the true state of the world k, and chooses how to report the state of the world k^* to the receiver. The receiver can either accept k^* as the true state of the world, or challenge the veracity of the report by calling *BS*. Senders are motivated to report an alternative state of the world k^* that advantages them most while still being believed by others. Thus they are constrained by two conflicting goals: (1) gain—a bigger lie (larger k^*) yields more rewards if accepted, and (2) plausibility—bigger lies are less plausible and more likely to be detected. Meanwhile, the goals of receivers are (1) to successfully detect lies to not be swindled, but (2) to avoid false accusations.

Receivers Detecting Lies

A receiver decides whether to call BS on a reported signal by computing the expected value of making an accusation for the reported signal, and comparing it to the expected value of accepting it as the truth. This calculation relies on combining the receiver's utility for calling BS (would I benefit from calling this message out as a lie?), with the receiver's beliefs about whether a given signal reflects the true state of the world (how likely is this message to be a lie?). This posterior belief arises from what the receiver believes of the sender's likely actions. The receiver must consider both what they believe the sender would report in each world state $P_S(k^* \mid k)$ and the distribution of true states of the world $P(k)$:

$$\mathrm{EV}_R(BS \mid k^*) \propto \sum_k U_R(BS; k^*, k) P_S(k^* \mid k) P(k) \tag{1}$$

Thus, choosing whether or not to call BS in response to a given report requires an estimate of how the sender decides what to report.

Senders Designing Lies

A sender decides what to report by calculating the expected value of each possible message based on their rewards and the likelihood that the receiver will call out a reported signal as a lie $P_R(BS \mid k^*)$:

$$\mathrm{EV}_S(k^* \mid k) \propto \sum_{BS} U_S(k^* \mid BS, k) P_R(BS \mid k^*) \tag{2}$$

Thus, the sender chooses not only whether to lie, but which lie to tell—potentially more rewarding but more conspicuous—based on their beliefs about how the receiver will respond to each message.

Equations (1) and (2) compute the expected value of the receiver's and sender's potential decisions, respectively. Both agents are assigned a probability that they will choose their actions by employing a Luce choice rule (Luce, 1959) over their expected values for each potential decision. In this way, the model builds in an assumption that receivers and senders rationally simulate the outcomes of alternative actions when deciding how to act.

Recursive Theory of Mind

If lying is a fundamentally dyadic, theory of mind reasoning problem, senders should lie and receivers should detect lies based on beliefs about their opponent's mental states and how they predict the other agent will make decisions. This means that equations (1) and (2) feed into each other: a receiver's decision to call BS is a function of their belief about the sender's actions; a sender's decision to lie is a function of their belief about the receiver's actions.

Whether a receiver calls BS and what a sender reports are defined via mutual recursion. Such recursive definitions might yield infinite computational complexity (if they were rolled out to infinite depth). In cooperative communication settings, mutual recursion converges given the concordant goals of the agents (Frank & Goodman, 2012; Schelling, 1960); however in adversarial settings, such recursion often fails to converge, and instead might cycle. We follow a conventional approach to resolve such non-convergent behaviors and follow the cognitive hierarchy model (Camerer et al., 2004) to define the agents as believing in a Poisson distribution over the depth of recursion that their opponent will consider. In other words, the sender may assume that their opponent is sometimes a 0-step receiver (i.e. calls BS randomly), a 1-step receiver (i.e. calls BS assuming the sender thinks the receiver is random), or an n-step receiver. However, rather than committing to a single assumption about the receiver, the sender assumes that the receiver is a weighted combination of all these potential strategies. The player then reasons one step further, choosing the best action in response to this weighted evaluation of their opponent's likely behavior. The Poisson distribution over opponent recursion depths smooths out cycling behavior in adversarial settings, and yields convergent results (Camerer et al., 2004). It is worth noting that the Poisson rate parameter is usually tuned to yield behavior consistent with humans, but is not independently verified to accurately track the distribution of reasoning depths of ecologically representative opponents. We refer to this strategy as the `Recursive Theory of Mind (ToM)` account of deception.

Alternatives to Dyadic Reasoning in Lying

Lying Heuristics

Many accounts of dishonest behavior do not assume that it arises from a rational consideration of alternatives, but instead is driven by certain inflexible strategies: heuristics. According to these accounts, individuals restrain their lies following simple self-oriented rules (e.g. Abeler et al., 2019; Gerlach et al., 2019; Mazar et al., 2008). For example, the prominent self-concept maintenance account hypothesizes that people lie by satisfying a constraint to preserve a concept of themselves as moral agents (Mazar et al., 2008). Notably, such accounts posit that lies face constraints from the liar's own values, prior beliefs, and knowledge about the true state of the world. These heuristics form a compelling alternative account, especially to help explain why people avoid saying maximal lies even in settings that do not bear a risk of detection. If people avoid large lies in the same manner regardless of whether they are at risk of being detected, one appealing explanation might be that listeners do not play a role in how people design lies after all; rather, it can be explained by individuals' lying heuristics.

We instantiate versions of these verbal theories as parametric models. The `Equal Intrinsic Aversion Heuristic` account posits that everyone shares the same intrinsic aversion to producing overtly large lies that results in people lying by some small amount on top of the truth. The second model assumes that people can be classified as those that exclusively tell the truth and others that lie (Hurkens & Kartik, 2009; Levine, 2019; Serota et al., 2010), again by some amount on top of the truth (`Unequal Intrinsic Aversion Heuristic`).

These alternative theories make several critically different predictions from the `Recursive ToM` account. Critically, both of these accounts predict that the size of lies should depend only on what the individual believes to be true, which serves as an anchor from which they adjust slightly. Furthermore, individuals in these accounts are about equally tempted to lie regardless of what the truth is. Additionally, these alternative accounts do not predict that lying behavior should change depending on what lies would be plausible to the receiver from base-rate beliefs about the world, nor is the decision to lie driven by payoffs. If receivers *cannot* respond or senders are not concerned about how the audience responds, then senders

have no motivation to reason about, or adjust their behavior to the audience's beliefs or goals. It is worth noting that both the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic accounts are models of lying behavior and make no prediction about lie detecting behavior.

0^{th} Order Theory of Mind

Let us now consider the minimally different account that specifically lesions the theory of mind component of `Recursive ToM`. This account *does* consider payoff gains for larger lies, so it is a rational agent, that decides what to report based on relative expected utility. However, this alternative agent *does not* attribute sophisticated reasoning to the audience, like having beliefs or goals that drive behavior. Instead, senders assume that the best receivers can do is to behave randomly when detecting lies (first-order intentional system; Dennett, 2009). Such a heuristic assumption about the opponent is not unreasonable: after all, previous work finds that people are practically at chance when detecting lies from the majority of liars (Bond & DePaulo, 2006; Levine, 2010), especially without enough useful context (Blair et al., 2010).

Specifically, this sender believes that their opponent will randomly and uniformly call BS without considering the payoff structure or statistics of the world. This model is equivalent to Equation (3) if the sender assumes the receiver is wholly a 0^{th} order reasoner. We call this the `0th Order Theory of Mind` model because the sender is merely attributing a primitive behavioral strategy to their opponent. If the sender believes their opponent behaves randomly, the sender need not adjust their lies to the statistics of the world; they can get away with lies of the same extremeness in any context. Senders under these conditions should lie maximally all the time, when they think that what they say has no bearing on the risk of being detected. Thus both the heuristic and `0th Order ToM` models predict that the sender will produce lies in the same manner regardless of the statistics of the world. For predictions of lying behavior on each of these accounts, see Supplemental Materials.

Predictions

If people believe their opponent uses theory of mind to lie and detect lies, then a rational sender ought to choose how and when to lie conditioned on how they assume a

rational receiver ought to call BS. Similarly, the rational receiver ought to call BS conditioned on how they assume the sender ought to lie or tell the truth—which will depend on the sender's beliefs about what the receiver will detect. Thus, both agents select their behavior conditioned on what a rational, albeit noisy, utility-seeking opponent would do, which results in a recursive process of reasoning about the other agent's likely actions. The `Recursive ToM` account generates four key predictions about lying and lie detecting behavior.

First, when the truth is less favorable, people should lie more often. Conversely, when the truth is already favorable, people have less motivation to lie, so they are more likely to tell the truth. Formally, this intuitive prediction arises out of the sender's goal to select actions that optimize their payoffs. Senders that do not consider payoff gains for producing larger lies will lie about equally often regardless of what was the truth, as in the `Equal Intrinsic Aversion Heuristic` and `Unequal Intrinsic Aversion Heuristic` accounts.

Second, people should balance payoff gains and plausibility when selecting what lie to send. While a larger lie might produce greater gains, a lie too large and implausible will be readily detected. Furthermore, when there are changes in the world that affect what might seem plausible, people should also adapt their lies. A hallmark of the `Recursive ToM` model is that it predicts rational senders should be attuned to prior beliefs and to the statistics of the world (i.e. the base-rate probability of an event or outcome). Of the models we compare, the `Recursive ToM` model uniquely predicts that senders' lies should be sensitive to base-rate information. Meanwhile, the `Equal Intrinsic Aversion Heuristic`, `Unequal Intrinsic Aversion Heuristic`, and 0^{th} `Order ToM` models predict insensitivity to the base-rate.

Third, plausibility is subjective — so, people should cater their lies to what is plausible to the specific audience in mind. What might seem plausible to a gullible audience may be scrutinized by a more knowledgeable audience, so people should hedge their lies accordingly. Showing that people's lies are sensitive to the specific and unique beliefs of their audience would be strong diagnostic evidence for a role of theory of mind in lying.

These predictions up until now have been about the sender's behavior that follows from considering the sender's goals and reasoning about what lies might be detected by the

audience. Are these valid assumptions about the audience's reasoning? As such, a fourth prediction is that these audiences are indeed sensitive to plausibility and payoffs. Receivers should robustly adjust their degree of suspicion based on changes to what seems plausible in the world. Alternatively, receivers may simply accept all reports as true, or randomly and uniformly call BS, ignoring goals and the plausibility of reports. Receivers may also be ignorant to just the payoff structure, in which case they should make neutral assumptions about the goals of the players. Instead, the best the receiver could do is to call out reports that are suspicious simply because they are unlikely to occur by chance. This process is akin to `Null Hypothesis Significance Testing`, in which the receiver has no bias to prefer lies of a certain direction. Lastly, receivers that ignore plausibility should not adjust how they call BS when the base-rate probability of an outcome changes.

Experiment 1

The core predictions of a rational, theory of mind based model of lying and lie detection are therefore: (1) People lie more when the truth is less favorable for them, (2) People craft lies by considering their plausibility and reward, and (3) People identify claims as lies based on their plausibility and payoffs. In Experiment 1, we test these predictions in a novel, dyadic lying game (Fig. 1) where people take turns reporting the number of red marbles drawn from a box, and classifying such reports as truths or lies. To test the core predictions of the `Recursive ToM` model against the predictions of alternative accounts, we manipulate the base-rate of red marbles in the box, and the payoffs associated with marbles of each color. Critically, we set up the incentive structure for senders and receivers so that senders are motivated to tell the biggest lie they can get away with, and receivers are motivated to call out lies, while avoiding false accusations. While these incentives may not reflect all real world lying situations, they capture common tradeoffs, and allow us to isolate the role of theory of mind in lying and lie-detection behavior.

Methods

Participants

A total of 228 participants were recruited from the undergraduate population at the University of California, San Diego. Two participants were excluded for producing out-of-bounds responses. Additionally, 14 participants were excluded for failing to meet the attention check criterion, which entailed achieving (within an absolute error of one) at least 75% (9 out of 12) numeric-response comprehension questions distributed throughout the task. After exclusion, 212 participants were included in the final data set. Participants were randomly assigned approximately evenly across the conditions (see Procedure). The study was conducted online, and participants were rewarded class credit for their participation. Informed consent was obtained from all participants, and all studies were approved by the university's Institutional Review Board.

Procedure

Participants played in a dyadic lying game that rewarded participants for strategic production and detection of lies. In each round of the game, both players saw a box of red and blue marbles which had some base-rate probability of sampling a red marble. The sender randomly sampled 10 marbles, of which k were red and the remainder were blue. When prompted about how many red marbles they sampled, the sender reported a number k^* which could be true or false. The receiver then saw how many red marbles the sender reported (with no knowledge of the true number) and could either accept this report or reject it as a lie.

Crucially, if the sender's report was accepted, the sender gained points for the number of red marbles reported k^* and the receiver gained points corresponding to the blue marbles reported $10 - k^*$ (in the condition where senders get points for red). So senders were motivated to lie and report an inflated value. However, if the report was rejected and the sender was indeed lying, the sender would lose points and the receiver would gain points. If the receiver rejected the report but the sender was in fact telling the truth, the receiver would face a penalty for making a false accusation, while the sender would receive points as they reported (Fig. 2 for full payoff structure).

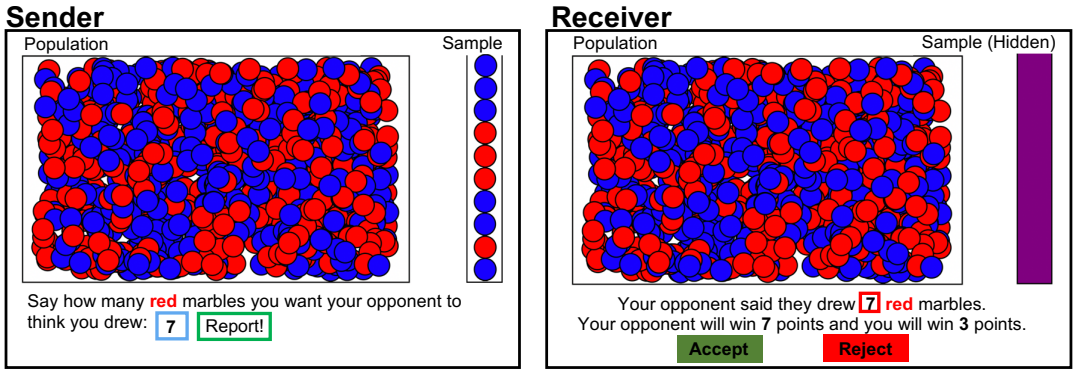

Figure 1

Experiment 1 used a dyadic lying game. The sender and receiver (one of whom is an AI; roles alternate across trials) both see the population of red and blue marbles (in the box; here, 50% red), but only the sender sees the true sample of 10 marbles (in the tube). (Left) Senders report the number of red marbles they sampled; they can tell the truth or lie by reporting something false. In this example, the sender gets points for red, while the receiver gets points for blue. The sender lies by reporting 7 red marbles, when in fact the sender actually sampled 4. (Right) Receivers accept or reject the reported number (i.e. call BS). If the receiver accepts, the sender gets 7 points for the reported red marbles, and the receiver gets 3 points for the reported blue marbles. If the receiver rejects, the sender gets caught in a lie and is penalized, while the receiver is rewarded.

In a 3×2 design, we manipulated (between-subjects) the base-rate probability of drawing a red marble (20%, 50%, 80%) and which color the sender got points for (payoff condition, *red* or *blue*) (Fig. 2). Across the payoff conditions, the mapping of points was reversed. When the sender got points for blue marbles (and the receiver for red marbles), the sender still reported the number of red marbles, so the sender was motivated to report *deflated* values (i.e. fewer red marbles corresponds to more blue marbles).

Participants played for 100 trials, switching roles between every trial, against who they were led to believe was another person but was in fact a computer. Participants were given the goal to win by the highest point difference possible. To discourage participants from learning from the computer's behavior, participants were not given feedback about their opponents' choice (i.e. the true number drawn, whether they lied or told the truth, accepted or rejected

report), after the initial practice trials. However, to motivate participants to pay attention, they were given updates on both players' cumulative points after every fifth trial. We expected that feedback only about cumulative points every fifth trial would not allow participants to learn or change strategy over time within the task. In line with this, we find that participants showed no performance improvement as the receiver, and only slight improvement as the sender—amounting to a $+0.7$ score improvement over 100 trials, and that could be attributed to an increased familiarity with the task (see Supplemental Materials for learning analysis).

Additionally, participants were intermittently asked trial-related attention check

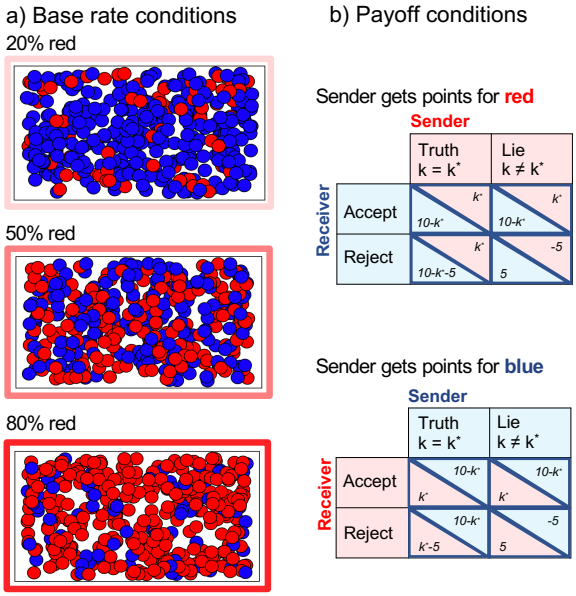

Figure 2

Experimental design. (a) Three base-rate conditions: the probability of sampling red marbles is 20%, 50%, or 80%. (b) Two payoff conditions: the sender gets points for red or blue marbles. Values in each triangular cell of the payoff table shows the points rewarded to each player (sender: top right triangle; receiver: bottom left triangle). Senders always reported red marbles k^*. Thus, when the sender gets points for red, the sender is motivated to report a higher number (more red), and when they get points for blue, a lower number (fewer red, therefore more blue). If the receiver catches the sender in a lie, the receiver is rewarded 5 points and the sender loses 5 points; if the receiver makes a false accusation, the receiver faces a -5 penalty atop what they would have received.

questions about how many red marbles they drew (if they just played as the sender) or the other player reported (if they just played as the receiver). Participants entered in a textbox their numeric response, and their possible responses were restricted to being between 0 and 10. The questions (12 in total) were randomly distributed after trials throughout the experiment (both practice and test trials).

Results

Senders' Lying Behavior

The behavior of senders can be divided into (a) their rate of lying (as opposed to truth-telling) and (b) the lie they told when choosing to lie. As we cannot pinpoint participants' underlying intentions, here we included lies as any reported value that was false, regardless of its intention. Reports grouped into this category may have been intentional lies designed to advance the player in the game, accidental false reports, etc. We compute the rate of lying (a) as the proportion of false reports to all reports. The lie told (b) is the report itself, conditioned on the report being false (and thus on the true number of red marbles sampled).

Senders lie more when the truth is less favorable to them. If senders lie more when reality is less favorable to them, we would expect the rate of lying to change as a function of how many red marbles they actually saw, such that the rate of lying increases with the number of red marbles seen when senders are rewarded for blue marbles, and to decrease with the number of red marbles seen when they are rewarded for red marbles.

We used the true drawn k, the payoff condition, and their interaction as predictors for the rate of lying in a mixed-effect logistic regression, with subject as a random intercept (Fig. 3). The payoff structure was treated as a sum coded factor. Critically, when senders got points for red, there was a negative slope of -0.28 ($SEM = 0.02$, $z = -17.14$, $p < 0.0001$), showing that people decreased their rate of lying when the true k was larger. In contrast, when senders got points for blue, there was a positive slope of $+0.15$ ($SEM = 0.02$, $z = 9.36$, $p < 0.0001$), so people increased their lying rate with larger k. Together, these results showed that people, guided by their payoffs, lie more when the truth is less favorable to them.

Senders lie by considering the plausibility and payoff of lies. The Recursive ToM

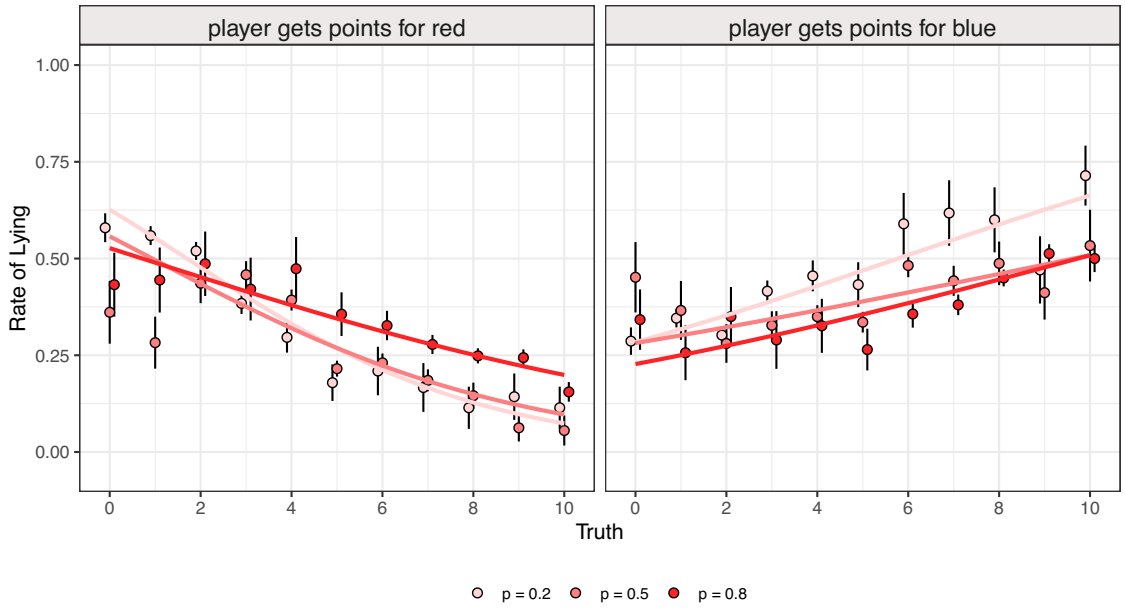

Figure 3

The rate of lying given the true sample for each condition in Experiment 1. Each point represents the rate of lying at a given truth value—the true number of red marbles sampled—by condition, and the error bars represent the standard errors of the mean. When senders get points for red, the rate of lying decreases as the truth increases; and vice versa for when senders get points for blue. These results show that people lie more frequently when the truth is less favorable.

model predicts that senders calibrate the extremeness of their lies to ambient base-rates (the probability of that outcome in the world). If the prevalence of red marbles in the box decreases, the receiver should be more suspicious about higher reported values, and therefore the sender should hedge by reporting fewer red marbles when they lie. Thus, under the `Recursive ToM` model, we would predict on average reported lies would become greater as the base-rate for drawing red marbles increases. In contrast, the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic models, and the 0^{th} `Order ToM` model, all predict no change in behavior as a function of the base-rate.

To test these predictions, we examined how the relationship between the true drawn k and reported lies (i.e. reported red marbles k^* when they differed from the truth) varied across the base-rate and payoff conditions (Fig. 4). We fit a linear regression to the number of

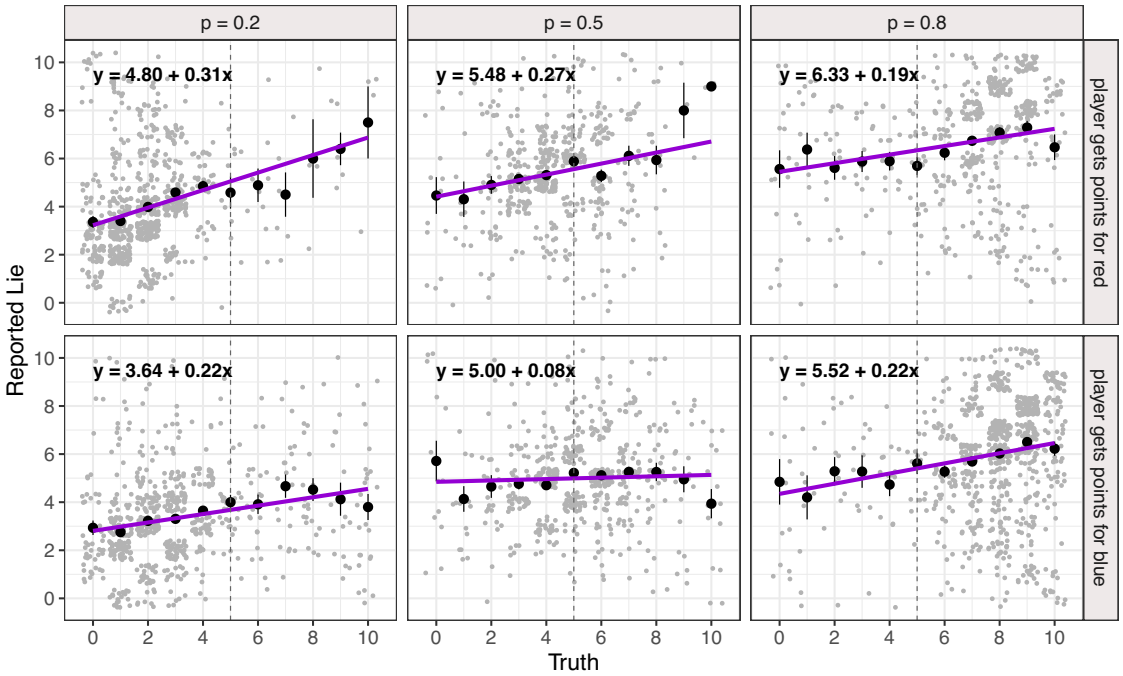

Figure 4

The distribution of lies from Experiment 1 participants across each condition. Each gray point was a false reported value. A linear mixed effect model was fit to each condition, with intercepts centered at Truth = 5. Intercepts increased across higher base-rate conditions, and there was a general shift across payoff conditions (top row vs. bottom row). These intermediary results allowed us to interpret differences in lies told across conditions and compare them to the model predictions in Fig. 5.

marbles falsely reported (i.e. k^* when $k^* \neq k$) with the predictors of the true value of k, the base-rate, the payoff structure, and the full interaction between these three factors. Subject was included as a random intercept. To facilitate comparisons across conditions, the true values of k were centered on 5 so that the models' intercepts correspond to the lies told when 5 marbles were truly drawn. Thus, changes in the intercept reflect changes in which lies are likely to be told in response to seeing 5 red marbles actually drawn.

First, we examined the general relationship between what lies the sender reported and what they actually drew (as the sole predictor in a fixed effect model). As expected, people's falsely reported numbers were larger when they drew more marbles in reality ($\hat{\beta} = 0.33$, $t(3865) = 27.17$, $p < 0.0001$, $r = 0.40$)[1]. This was true regardless of whether someone is

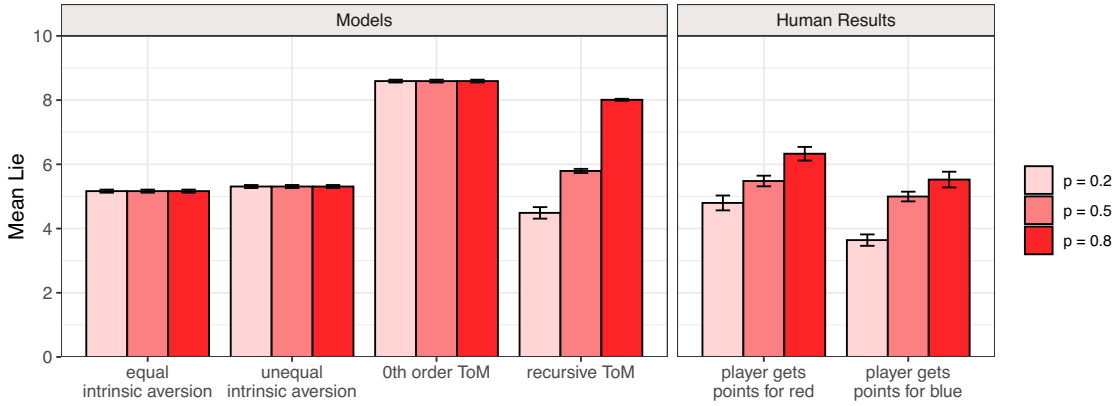

Figure 5

The model prediction and human results for the mean lie, computed from the intercept of the linear fit (e.g. from Fig. 4). The Recursive ToM model uniquely predicts that the sender should alter their mean lie based on the receiver's base-rate belief. Results from Experiment 1 show that human participants calibrated their lies to the base-rate for sampling red marbles.

motivated to lie by over-reporting (as in the red payoff condition; $\hat{\beta} = 0.41$, $t(1779) = 22.42$, $p < 0.0001$, $r = 0.47$) or under-reporting (blue payoff; $\hat{\beta} = 0.33$, $t(2084) = 19.80$, $p < 0.0001$, $r = 0.40$).

To address our main question, we next analyzed whether the base-rate condition influenced people's lies. Critically, the sender's reported lie significantly changed across the base-rate conditions ($\chi^2(8) = 87.8$, $p < 0.0001$). When senders got points for red marbles,

[1] When thinking about the magnitude of people's lies (the distance away from the truth) as a function of the true value, it is important to note that the restricted reporting range of the task means that the magnitude of possible lies is more restricted toward the ends of the range. For example, if the goal of the sender is to over-report how many red marbles they saw, then when fewer red marbles are sampled, there is a greater margin for over-reporting. In this case, people cannot possibly lie by the same magnitude when they see a large number as compared to when they see a small number. In Fig. 4, a slope of 1 would indicate a constant difference between truth and lies regardless of how many red marbles were actually drawn. A slope of 1 for these task results would be impossible unless the average lie magnitude was 0—when the truth was 10, speakers cannot possibly lie in the positive direction, since they can only report numbers between 0 and 10. Our results showed a much shallower slope of 0.33, revealing that the magnitude of the lie was smaller for larger truths. However, this behavior may have arisen from the restricted range of the task.

their lies were highest when the base-rate was 80% (Mean lie = 6.33, SEM = 0.21), intermediary when the base-rate was 50% (Mean lie = 5.48, SEM = 0.17), and lowest when the base-rate was 20% (Mean lie = 4.80, SEM = 0.23). These results are in line with the predictions of the `Recursive ToM` model—that senders calibrate their lies by reasoning about what the receiver may find plausible from base-rate information about the world. The payoff condition (red vs. blue) was also a significant predictor of reported lies ($\chi^2(6) = 32.0$, $p < 0.0001$). In other words, the sender's goal additively shifted the sender's reported lies. This means that in the blue payoff condition the mean lie was also larger for higher base-rate conditions (of red marbles), except with an overall drop in magnitude relative to the red payoff condition. When the sender got points for blue marbles, the mean lie at a base-rate of 80% was 5.52 (SEM = 0.25), 5.00 when the base-rate was 50% (SEM = 0.15), and 3.64 when the base-rate was 20% (SEM = 0.18). Thus both the base-rate and payoff conditions additively affected what lies people reported.

Our evaluation of the sender's lying behavior confirmed the core predictions of the rational, theory of mind based lying model and violated the predictions of the alternative heuristic models (Fig. 5). Namely that senders lie more when the truth is less favorable to them, and they choose lies by considering both their plausibility and the players' payoffs.

Receivers' Lie Detecting Behavior

The `Recursive ToM` model of rational, statistical senders assumes that receivers are themselves rational, statistical agents. Specifically, it assumes that receivers are more likely to identify a claim as a lie if it is less plausible and more consistent with the reward structure. However, the prevailing view is that human lie detection behavior is close to chance (54% accuracy; Bond and DePaulo, 2006; Gladwell, 2019; Levine, 2014). If receivers are at chance, random, or otherwise insensitive to how a claim compares to the relevant statistical information in the world, then reports should not be called out based on simple base-rates. Alternatively, the receiver may have preferences that are not dependent on the relevant goals. In this case, the report least likely to be called out should simply be the most likely number arising from random sampling (e.g. 5 red marbles, when 50% of the marbles in the population are red). This is the prediction of the `Null Hypothesis Significance Testing (null)`

account. Within each of these accounts, receivers do not exhibit the sophistication we attribute to the senders' model of the receiver. Alternatively, if the receiver prefers reports of fewer red marbles, or if the receiver knows the sender is motivated to report more red marbles, then reports of more red marbles are more likely to be called out, and the reports that the receiver accepts as true will have fewer red marbles on average. Do real human receivers detect lies in the rational manner we have assumed in our senders' model?

Figure 6ab shows the rate at which receivers reject a given reported number of marbles, revealing that receivers are more likely to reject as a lie reports of many red marbles when the base-rate of red marbles is lower (indicating a sensitivity to the plausibility of reports), and when red marbles are rewarded for the sender (indicating a sensitivity to payoffs). To quantify these patterns, we characterized the receiver's behavior in terms of the report that they were most likely to accept (i.e. least likely to call out as a lie) in Figure 6c. We estimated this value by taking the maximum-likelihood of a (vertex-form) quadratic logistic regression fitted to the human receiver data. This allowed us to infer the report for which receivers least called BS for each condition (see Supplementary Materials for more details).

Receivers find lies that are more consistent with the base-rate to be more plausible. We found that human receivers adjust which reports they call out based on the base-rate probability of sampling red marbles (Fig. 6). Collapsing over payoff conditions, in the 20% base-rate condition, the mean most accepted report was a report of 3.94 ± 0.22 red marbles. The mean accepted report was larger for higher base-rate conditions, with a report of 5.13 ± 0.17 in the 50% base-rate condition and 5.82 ± 0.24 in the 80% base-rate condition. The pairwise differences across all conditions were significant: at 20% vs. 50% ($z = 4.30$, $p < 0.0001$), 50% vs. 80% ($z = 2.40$, $p < 0.02$), and 20% vs. 80% ($z = 5.83$, $p < 0.0001$). This shows that receivers detect lies based on their consistency with statistical information they believe about the world.

Receivers are more likely to identify claims as lies when they are aligned with the reward structure. The human receivers' mode of accepted reports also differed depending on the payoff structure. When receivers were rewarded for blue marbles (rather than red) they tended to accept reports with more blue marbles, for all base-rate conditions. In the 20%

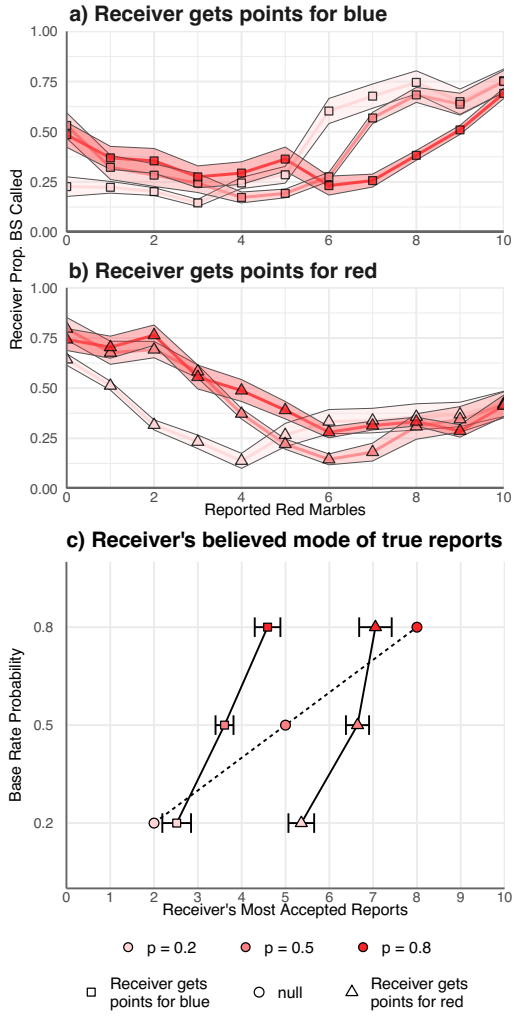

Figure 6

Human results for the receiver's rate of calling BS (rejecting the reported value as a lie) based on how many red marbles the sender reported. As predicted by `Recursive ToM`*, (a) when receivers got points for blue, receivers more often called out reports of high numbers of red marbles as lies; (b) when they got points for red, the opposite was true. Receivers accounted for the statistics of the world in detecting lies, shown by the shift in BS-calling across base-rate conditions. (c) We summarized the results of (a, b) by estimating which reported value receivers are most likely to accept. This quantity can be interpreted as the implied mode of the believed true reports (x-axis). The error bars represent the 95% confidence interval of the mean. Under the null—receivers are ignorant to the payoff structure—the mode would be equal to* $10\times$ *the base-rate and would not vary by payoff condition. Instead, receivers' behavior varied systematically with the payoff condition.*

base-rate condition, receivers' most accepted report was 2.52, 95% CI = [2.20, 2.84] red marbles when receivers were rewarded for blue marbles, and 5.36, CI = [5.07, 5.65] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 2.84$, $z = 12.93$, $p < 0.0001$). In the 50% base-rate condition, the receivers' most accepted report was 3.61, CI = [3.41, 3.81] red when receivers were rewarded for blue marbles, and 6.64, CI = [6.38, 6.90] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 3.04$, $z = 18.15$, $p < 0.0001$). Lastly, in the 80% base-rate condition, the receivers' most accepted report was 4.59, CI = [4.30, 4.88] red marbles when receivers were rewarded for blue marbles, and 7.05, CI = [6.69, 7.42] red marbles when receivers were rewarded for red marbles ($\bar{x}_d = 2.46$, $z = 10.41$, $p < 0.0001$; Fig. 6c). These results conclude that receivers call out lies by considering their alignment with the reward structure for both players.

Discussion

Experiment 1 tested several predictions of the `Recursive ToM` model in a dyadic lying game where players took turns reporting to an adversary the number of red marbles drawn from a box, and classifying their adversary's reports as truths or lies. Critically, we manipulated the base-rate of red (vs. blue) marbles, as well as the payoffs associated with more red marbles for both players. We found support for three key predictions of the rational, theory of mind based model of lying and lie detection, namely that: (a) people lie more when the truth is less in their favor, (b) people choose lies based on their plausibility and payoffs, and (c) that people are also sensitive to plausibility and payoffs when detecting lies. In contrast, we found that lying behavior did not fit with the predictions of several alternative models.

While the `Equal Intrinsic Aversion` and `Unequal Intrinsic Aversion` heuristic models predict that people may ignore external payoff gains, and so should lie equally often regardless of the truth; people, in fact, do change how often they lie as a factor of the truth. Additionally, while both the heuristic models and the 0^{th} `Order ToM` model predict that the lies people do say will be insensitive to the base-rate; people, in fact, tune their lies based on what lies could be plausible. Lastly, while the `Null Hypothesis Significance`

`Testing` model predicts that receivers, having no bias to prefer lies in a certain direction, should call out large and small reports equally often; people, in fact, consider the payoff structure when deciding how to call out lies. Under these considerations, the `Recursive ToM` models seems to accurately predict human lying and lie detecting behavior.

However, this experimental design cannot test a more subtle claim of the `Recursive ToM` model: that people tailor their lies to the beliefs they attribute to the receiver, even when those beliefs are different from their own. We test this claim in Experiment 2.

Experiment 2

Experiment 2 tested how people lie when senders and receivers have divergent beliefs about the probability of the world. We expand on the design of the lying game from Experiment 1: Now the distribution of marbles is only partially observable to the receiver, limiting their visual access. Importantly, the sender can observe what the receiver sees, but also has greater visual access, sometimes resulting in the sender having different beliefs about the base-rate than the receiver. This design serves to tease apart whether senders simply adjust their lies to *their own* beliefs or to beliefs about *their audience's* beliefs. Critically, a lying strategy that calls upon theory of mind to avoid detection ought to adapt lies to the audience's beliefs, and not simply to only one's own beliefs, in contrast with all accounts that generate lies without considering the listener. Thus, in Experiment 2, we address this question by dissociating the sender's and receiver's base-rate beliefs to investigate whether and how the receiver's beliefs influence a sender's lies.

Methods

Participants

291 participants were recruited from the undergraduate population at UCSD. Of these, 33 were excluded for failing to sufficiently answer at least 75% of the attention check questions. Therefore, 258 participants were included in our final data set.

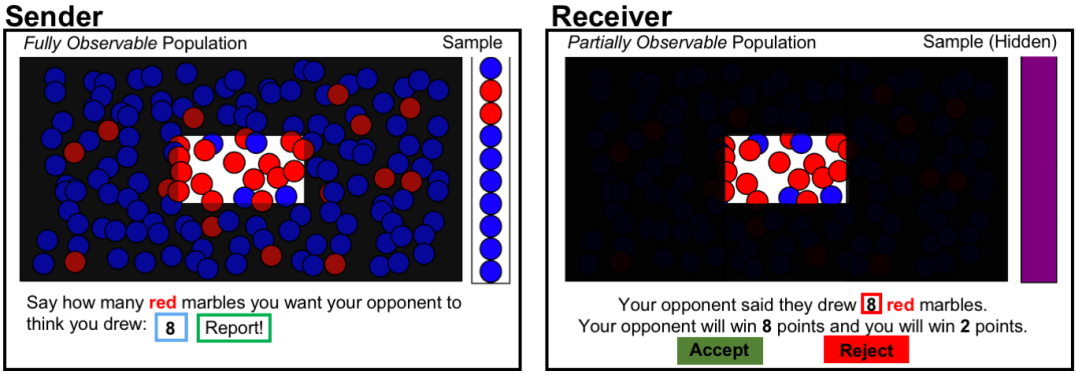

Figure 7

Experiment 2 used a partially observable dyadic lying game. For the sender, beliefs about the base-rate are fully observable: the sender knows the distribution of red and blue marbles observed by the receiver (in the inner white box), and the overall distribution (in the inner white and surrounding black box). For the receiver, beliefs about the base-rate are partially observable: the receiver can only observe the distribution of marbles in the window (the inner white box). Here, the sender believes the full population contains 20% red and 80% blue marbles, and they know the receiver observes a subset of the population that is 80% red and 20% blue marbles.

Procedure

The lying game used in Experiment 2 resembled the game introduced in Experiment 1, except it separately manipulated the distribution of red and blue marbles in the box visible to the sender and the receiver (Fig. 7). In Experiment 1, the box of marbles was fully visible to both players; in Experiment 2, the box contained a window on one side (an inner white box) through which the receiver could see the distribution of red and blue marbles. The other side was open—the sender could see what the receiver saw through the window (the inner white box), as well as the full distribution of red and blue marbles (the inner white box and the surrounding black box). In other words, the population of marbles was fully observable for the sender, but only partially observable for the receiver. The sender could infer how the receiver's base-rate differed from their own, but the receiver had no information on which to evaluate whether the sender had a belief different from their own. As in Experiment 1, participants alternated between playing as the sender and receiver.

We used a 3×3 within-subject design manipulating: the sender's base-rate (total box; 20%, 50%, or 80% red); the receiver's base-rate (inner white box; 20%, 50%, or 80% red). This within-subject design necessarily required more participants, relative to Experiment 1. These conditions were randomly sampled for each trial.

To check whether our manipulation resulted in senders and receivers having divergent beliefs (as we intended), we asked participants to respond on a slider scale about the distribution of marbles from their own or their opponent's perspective (shown in Fig. 8). The left side of the slider bar was red and the right side was blue, so that the further rightward the bar was dragged, the more the bar was "filled in red." Labels below the slider ("more blue" to the left, "more red" to the right) helped to clarify the scale's direction. As in Experiment 1, participants also answered randomly distributed attention check questions about the number of red marbles drawn or reported. All participants received a total of 19 base-rate and attention check questions, except three subjects who received 18 (due to randomization).

In addition, unlike in Experiment 1 which manipulated the payoff structure of the game across conditions, Experiment 2 used only the red payoff condition's utility structure. In other words, senders generally received points for more red marbles (and were motivated to over-report what they saw), and receivers received points for more blue marbles. Once again, participants played for a total of 100 trials.

Results

Manipulation Check

Did our manipulation of sender's beliefs, receiver's beliefs, and sender's beliefs about the receiver have the intended effects? For our manipulation to work, three conditions must be satisfied: (1) The sender must recognize that the receiver only has visual access to the distribution of marbles in the inner white box, and it guides the receiver's beliefs. (2) The sender must recognize that each player can hold different beliefs about the base-rate of marbles. (3) The receiver must actually believe the base-rate that they see, so as to make them susceptible to exploitation. We evaluated if participants' base-rate estimates (ranging from 0 to 100) varied as expected with player role (sender or receiver), question type (own or

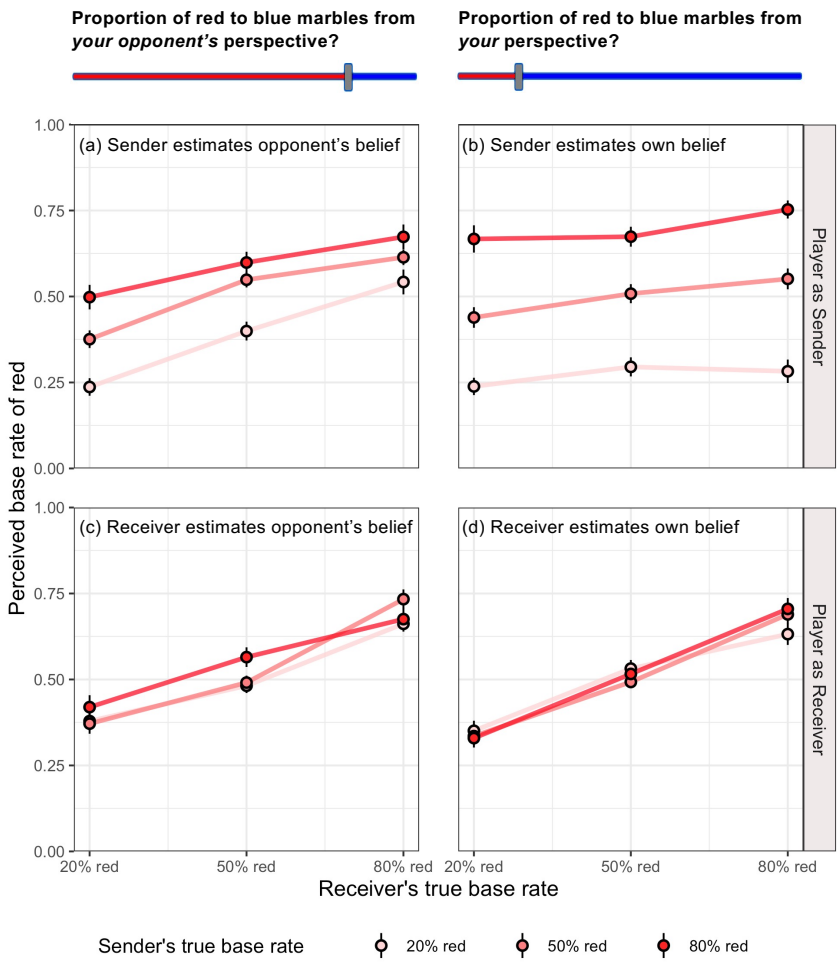

Figure 8

Participants reported their beliefs about the distribution of red/blue marbles. The x-axis is the receivers' base-rate condition, and the color is the senders' base-rate condition. The panel rows indicate the role of the participant as the sender (top) or receiver (bottom), and the panel columns indicate if the participant was asked about their opponent's (left) or their own beliefs (right). The y-axis, shows the participants' slider scale response. (a) When senders estimated their opponent's (receivers') beliefs, senders believed receivers' base-rate beliefs shifted with the receivers' true base-rate as expected, but surprisingly, the senders' true base-rate also had a small influence on their response. (b) When senders estimated their own (senders') beliefs, senders accurately assessed their own base-rate. (c) When receivers estimated their opponent's (senders') beliefs, and (d) when receivers estimated their own (receivers') beliefs, receivers responded the same.

opponent's belief), and sender and receiver base-rate conditions.

Does the sender notice the receiver's base-rate? We checked if the study's key manipulation was successful—that the sender was aware of the receiver's beliefs, informed by the inner white box (Fig. 8a). A two-way ANOVA with an interaction revealed a significant effect of receiver base-rate ($F(6,636) = 17.06$, $p < 0.0001$), suggesting that sender understood that the receiver's beliefs about the base-rate were constrained by the aperture. There was also a significant effect of sender base-rate ($F(6,636) = 11.59$, $p < 0.0001$), indicating some "leakage" of the sender's beliefs into their assessment of the receiver's beliefs.

Does the sender believe the receiver has divergent beliefs? Our manipulations were specifically aimed to induce an asymmetry between the sender's beliefs about the receiver's beliefs (Fig. 8a) and the sender's own beliefs (Fig. 8b). We tested whether *whose* beliefs (sender or receiver) the sender was asked about interacted with the receiver and sender base-rate conditions, separately. For both interactions we found a significant effect (with receiver base-rate: $F(2,1242) = 10.72$, $p < 0.0001$; with sender base-rate: $F(2,1242) = 21.74$, $p < 0.0001$). This means that our manipulations succeeded at separately influencing the sender's estimates of the base-rate, and their assessments of the receiver's beliefs about that base-rate.

Does the receiver assume the sender shares the same beliefs as themselves? Another assumption of the study is that receivers assume that the distribution of marbles visible to them approximately matches the distribution of marbles from which the sender is sampling. Alternatively, as the receiver, the participant may believe the game is rigged and distrust that the distribution for the players will match. To address this issues, we correlated the receiver's mean perceived base-rate beliefs about the sender's beliefs (Fig. 8c) and their own beliefs (Fig. 8d), and we found that the responses were positively correlated ($r = 0.74$, $t(7) = 2.94$, $p < 0.03$). This suggests that the receiver's belief about the sender's and their own beliefs approximately mapped onto each other for each condition, corroborating our assumption that the receiver defaults to assuming the sender's beliefs approximate their own.

In aggregate, our manipulations worked. The sender recognized that the receiver's beliefs about the base-rate were different from their own, and the receiver used their visible

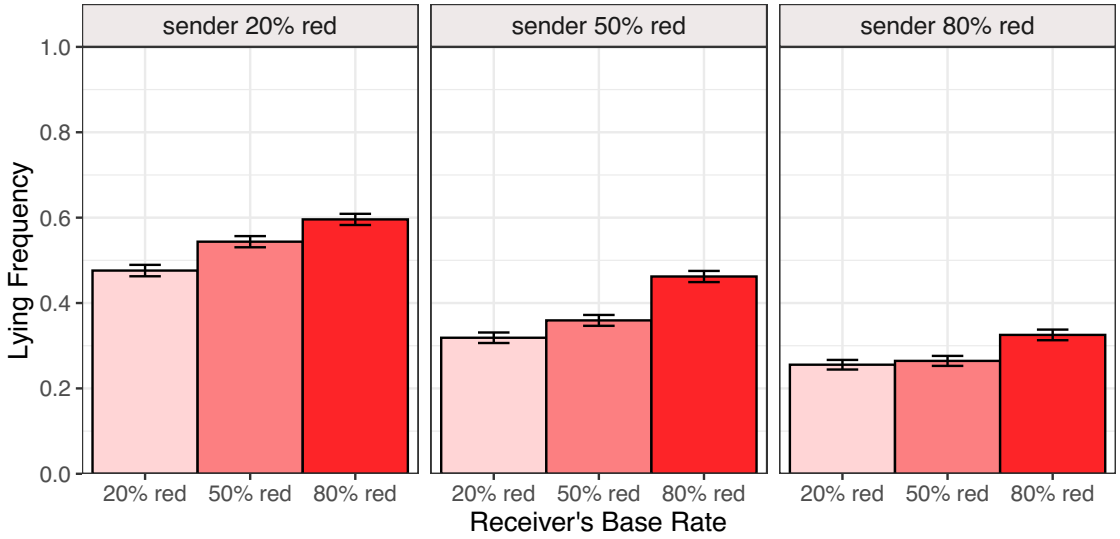

Figure 9

The rate of lying (as opposed to telling the truth) across conditions. There is as an effect of the receivers' (x) and the senders' base-rate condition (panels). People lie more when the receivers' base-rate belief is higher (e.g. 80%), suggesting that people recognize when their audience is more exploitable.

information to approximate the sender's likely beliefs.

Speakers design lies by considering receivers' prior beliefs

When do people lie? Under the `Recursive ToM` account, people should lie more when they believe the receiver has a higher base-rate belief, as having a high base-rate belief leaves people more susceptible to believing a large lie could be true (Fig. 9). In line with this, the sender's lying frequency increases with the true base-rate belief of the receiver ($\chi^2(2) = 138.5$, $p < 0.0001$), as well as the true base-rate experienced by the sender ($\chi^2(2) = 664.8$, $p < 0.0001$). These results imply that people can recognize when their audience is more exploitable, and they more frequently take advantage and lie in these situations.

How do people lie? Next, we examined how people chose to lie as a function of their own and the receiver's base-rate beliefs. As in Experiment 1, we extracted the centered intercept from the linear relationship between the true number of red marbles sampled and

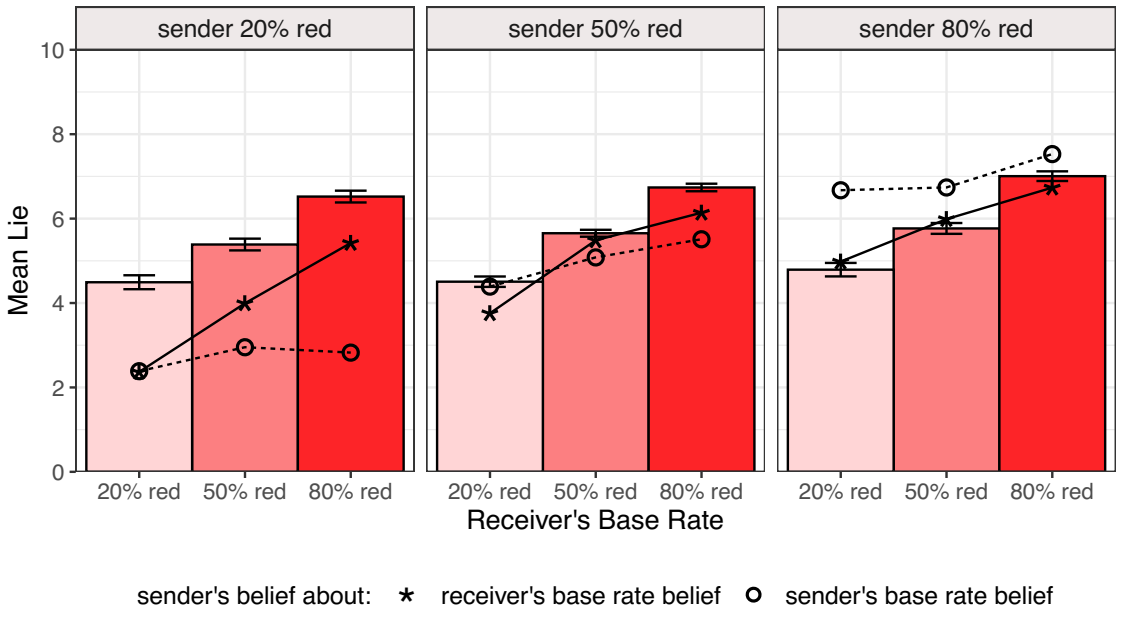

Figure 10

The average lie across conditions, computed from the intercept of the linear fit. There is a strong effect of the receivers' base-rate condition (x-axis), and little effect of the senders' base-rate condition (panels). Stars represent the senders' estimates of the receivers' belief about the base-rate (from Fig. 8a), and circles represent the sender's direct estimate of the base-rate (from Fig. 8b). The average lie more closely tracks senders' estimates of receivers' beliefs, suggesting that senders use theory of mind to choose how to lie.

reported lies for each sender base-rate and receiver base-rate conditions. We then used this value as a summary of the sender's average lie in order to examine whether senders' or receivers' base-rate beliefs influenced people's lies, and to compare which was the stronger predictor (Fig. 10). The results revealed that both of the base-rate conditions were significant predictors of the reported lies, but the receivers' base-rate had a greater effect on lies ($\chi^2(12) = 1214.7$, $p < 0.0001$, $\hat{\omega}^2 = 0.119$) than the senders' base-rate ($\chi^2(12) = 34.7$, $p < 0.001$, $\hat{\omega}^2 = 0.003$). Thus, senders weighed receivers' prior beliefs *more* than their own when deciding how to lie. These results point to people's abilities to construct gain-increasing lies around the audience's unique beliefs and support the claim that senders are using an audience-based strategy to choose their lies.

Discussion

Experiment 2 sought to tease apart whether a sender designs lies that are solely influenced by the sender's own beliefs, or their ToM-driven reasoning about the receiver's beliefs. The latter is qualitatively predicted by the `Recursive ToM` account of lying, which uniquely considers the beliefs of the receiver. Thus, the partially observable lying game allowed us to evaluate the role of ToM, by considering how people lie when there is an explicit mismatch between their own prior beliefs and their estimates of the receiver's prior beliefs. In these settings, we found that peoples' lies are better predicted by beliefs about the receiver, as opposed to beliefs about themselves, further supporting a critical role of theory of mind in deciding how to lie.

General Discussion

The current work presents a unified framework underlying lie design and detection, formalized as recursive social reasoning. This approach highlights how liars and lie detectors plan their behaviors via interactive, adversarial reasoning: Senders design lies by inferring the likelihood the receiver detects potential lies; receivers detect lies by inferring if and how the sender would lie. We compared our `Recursive Theory of Mind (ToM)` account to three accounts that do not require ToM, namely the `Equal Intrinsic Aversion Heuristic`, `Unequal Intrinsic Aversion Heuristic`, and 0^{th} `Order ToM` accounts. Compared to the other models, `Recursive ToM` uniquely generated several key diagnostic qualitative predictions about patterns in lying and lie detecting behavior: (1) people should lie more when the truth is less favorable, (2) people should balance payoff gains and plausibility when deciding what lie to say, (3) people should cater their lies to what they think their audience will find plausible, and (4) when detecting lies, people should be sensitive to plausibility and payoffs, as well.

We empirically tested and showed that our model explained the rate and content of lies people produced, and which lies they detected. In Experiment 1, people lied more when the truth was less favorable, consistent with lying being strategically scaled to circumstances, rather than being a small constant offset if lying was based simply on anchoring to speaker's

knowledge of the truth. Furthermore, people produced larger lies when those larger lies were more consistent with the base-rate, indicating that they balanced payoff and plausibility, in contrast with the idea that lies are tempered largely by moral considerations. Finally, people detected lies by being sensitive to payoff and plausibility, indicating that lie detectors are sensitive to the content of the lie, rather than solely considering superficial cues. Experiment 2 provides stronger diagnostic evidence for a role of theory of mind in lying: when senders and receivers have a mismatch in beliefs about the world, senders tuned their lies to the audience's beliefs more than to their own beliefs. This further confirms that lies are crafted to balance payoff and plausibility for the listener. Altogether, we take these results as evidence that people can and do spontaneously calibrate their lying and lie detecting by employing theory of mind.

The idea that people reason about others' minds when strategizing about lies builds on evidence from previous work. For instance, children's theory of mind development is linked to their ability to lie and to maintain their lies over time (Ding et al., 2015; K. Lee, 2013; Talwar et al., 2007). In adults, beliefs about the receiver's level of suspicion predicts whether or not people choose to deceive (Franke et al., 2020; Gneezy et al., 2018; Montague et al., 2011; Nagin & Pogarsky, 2003; Ransom et al., 2019; Rogers et al., 2017). However, the current results support a stronger claim: theory of mind underlies a unified model of lying and lie detection—the frequency *and* magnitude of lies are calibrated using people's interactive social reasoning. People rationally lie and detect lies, using social reasoning to predict other agents' behavior.

The view of lying and lie detection supported by our experiments—as strategic acts driven by theory of mind reasoning—adds nuance to the prevailing focus in the literature. Previous research on lying often considered high-stakes, hard to detect, lies—such as in tax evasion and fraud; or during police interrogations—and asked distinct questions, such as whether highlighting the moral salience of lying could increase honesty (Kristal et al., 2020; Mazar et al., 2008), or whether superficial behaviors could be used to detect lies (DePaulo et al., 2003). This work emphasized that lies can be constrained by intrinsic morals (Mazar et al., 2008), and that people are poor at performing lie detection in some circumstances (Bond

& DePaulo, 2006), in part because they are too automatically trusting to succeed (Levine, 2014). In contrast, here we focus on common, everyday lies—lies that are commonplace but where extreme lies are easily detectable, analogous to overstating your resume qualifications, or lying about your height on dating profiles. We show that theory of mind shapes the generation and detection of these common, everyday lies. While we agree that moral considerations (for example) can modulate lies, and that there are limits to people's lie detection abilities, the current work suggests that theory of mind plays a foundational role in lie generation and detection, acting in concert with these additional factors. We expect this is true of high-stakes lies as well.

In real-world settings, this strategic theory of mind strategy and a moral individually-focused lying aversion are not mutually exclusive. People may lie, for example, by primarily trading off maximizing their gain and avoiding audiences' detection, but they may secondarily avert conventionally unethical lies. Both cognitive mechanisms are likely weighed variably across contexts, which may partly explain why the propensity to lie varies across experimental paradigms and laboratory versus field studies (Gerlach et al., 2019). Our results indicate that human lying behavior is not driven solely by individual factors, but instead takes into account what others are likely to think about potential lies. However, this does not mean that there is no role for individual factors—at the very least, there is likely to be individual variation in aversion to lying, even though lies, when told, are strategically designed for the audience. Future work may more directly compare how other factors, like moral reasoning, trades off with audience-related factors.

Our results relate to the literature on recursive reasoning in behavioral game theory. Classic work in this domain classifies agents as level-*k* reasoners (Crawford & Iriberri, 2007; Stahl, 1993), or describes their reasoning capacity in terms of a cognitive hierarchy of recursion depths (Camerer et al., 2004). Work in this field has attempted to characterize peoples' exact recursion distribution, using games designed explicitly to measure the number of levels of recursion each person has computed—such as the p-beauty contest (Ho et al., 1998; Nagel, 1995). This work shows that people are well-characterized by an average recursion depth of 1.5 (Camerer et al., 2004). How many levels of recursion do people

compute in adversarial communication contexts, when lying or detecting lies? Our experiments demonstrate that in adversarial communication contexts, listeners and speakers are both at least level-2 reasoners: Listeners consider the goals of speakers when detecting lies, and speakers consider the beliefs of listeners when designing lies. This provides a lower bound on recursion depth of participants in our experiment. However, our data cannot establish an upper bound, or participants' precise *k*-level. Future work should adapt finer-grained methods to identify the level of recursion that people employ during adversarial communication.

We intentionally set up our experiments to resemble the common situation in which larger magnitude lies are both more rewarding if accepted, and also easier to detect, to create simple countervailing pressures on liars. This situation occurs for many real-world situations in which numbers are reported—fraud about balance sheets, taxable income, and revenues, for example. All have the property that lie magnitude monotonically increases reward (if believed), but also monotonically increases the risk of detection. However, detectability may not always trade off with value to the sender: For example, imagine the subtle yet highly advantageous lie possible when a test is graded pass/fail, and a students' score is only one percentage point from the threshold. In this case, fudging the score by only one point results in a change from no credit, to full credit. More generally, this situation arises when the utility (value) of an answer does not scale linearly with the possible responses. We expect that the same kind of recursive theory-of-mind based reasoning seen here would be used in this more complex case. In other words, when both the sender and receiver are aware of this non-linear distribution of utility, their lie detection should adjust to consider responses more likely to be lies when they are more in line with the senders' goals, even though this consideration is more complex to consider than simply larger numbers equating to higher utility. Future work may test whether people employ recursive social reasoning when faced with more complex, non-linear distributions of utility.

The current experimental setting also concerns a situation in which accurately detecting and calling out a lie is advantageous. While this is true in many real-world contexts, lying interactions are often more complex. For example, lie detection may be strategically

concealed. There are many settings where, upon detecting a lie, the most prudent action is to not reveal that the lie has been detected (perhaps due to some utility cost to revealing private information, or for instigating conflict). This strategic concealment adds another deception to the situation. In this situation, the receiver faces a decision: Whether to tell the truth and state that a lie was detected; or whether to lie by omission, and choose not to reveal that a lie was detected. More broadly, agents' decisions to show or hide knowledge about other agents' goals and beliefs likely plays a crucial role in the arms race between deceivers and detectors. Characterizing the reasoning underlying strategic concealment of lie detection is an important next step in expanding the scope of our model, toward understanding the natural complexity of adversarial communication.

Like all laboratory experiments, ours were designed to isolate and measure particular effects: In the research we present here, we show that people can rationally lie by considering their opponent's beliefs and goals. Our experiments used a low-stakes, game-like setting, potentially increasing subjects' willingness to lie, and to lie strategically, as compared to high-stakes real-world communication. Similarly, lying in our experiments was compartmentalized from subjects' real lives, thus eliminating considerations of reputation management and downstream reciprocity. By having participants play against an artificial agent, we can control how opponents behave, to more effectively measure how people respond; but surely incentives will differ in this setting compared to face-to-face communication. Furthermore, our results pertain to the behavior of many individuals, compared in between-subject conditions in Experiment 1, or aggregated in Experiment 2—in some cases group behavior may appear to fit a particular model, while individuals do not (Goodman et al., 2008; Vul et al., 2014). Future work should examine individual differences in reasoning, to evaluate the extent to which key recursive reasoning patterns apply to each participant considered in isolation.

Lastly, by alternating roles on consecutive game rounds, participants may have become more likely to consider how their counter-party would respond to their behavior, relative to situations in which they had never experienced being the counter-party. This would be in line with developmental evidence for the role of first-person experience in some aspects of action

understanding (e.g. Gerson & Woodward, 2014). Furthermore, repeatedly taking turns between roles seems to mirror the level-*n* recursive theory of mind reasoning described in our model. How might the act of alternating roles facilitate recursive reasoning? Future work should explore to what extent first-person experience as both the sender and receiver facilitates or is necessary for reasoning about others' beliefs, goals, and actions in the context of lying and lie detection. In the context of our task, this could involve having participants play only one role, or presenting roles in blocks of trials rather than in alternation.

Overall, we provide evidence that people can spontaneously calibrate their lying and lie detecting by employing theory of mind. Incorporating mental state reasoning into the understanding of lies may be a useful path toward smarter, more human-like AI to automate the detection of false information ("fake news") on social media. False claims are highly prevalent online, and motivated by particular goals; but current AI systems do not incorporate others' likely knowledge and motives, limiting lie detection. The current data suggest that a greater emphasis on socially intelligent mechanisms is warranted in our push towards more epistemically vigilant AI systems (Sperber et al., 2010). Overall, our work both advances basic science of cognition and mental state reasoning, and moves toward an automated system for improved detection of false claims.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153.

Allcot, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844.

Arrow, K. J. (1965). *Aspects of the theory of risk-bearing*. Helsinki, Yrjö Jahnssonin Säätiö.

Becker, G. S. (1968). Crime and punishment: An economic approach, In *The economic dimensions of crime*. London, Palgrave Macmillan.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research*, *36*(3), 423–442.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Bond, C. F., Howard, A. R., Hutchinson, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology*, *35*(2), 212–221.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515.

Bruer, K. C., Zanette, S., Ding, X. P., Lyon, T. D., & Lee, K. (2019). Identifying liars through automatic decoding of children's facial expressions. *Child Development*.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*(3), 203–242.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93–99.

Crawford, V. P., & Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, *75*(6), 1721–1770.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257.

Dennett, D. (2009). Intentional systems theory, In *The oxford handbook of philosophy of mind*.

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science*, *26*(11), 1812–1821.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*(1), 88–106.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*(3), 414–420.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723–733.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise–an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M., Dulcinati, G., & Pouscoulous, N. (2020). Strategies of deception: Under-informativity, uninformativity, and lies — misleading with different kinds of implicature. *Topics in Cognitive Science*, *12*(2), 583–607.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44.

Gerson, S. A., & Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child Development*, *85*(1), 264–277.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, *20*(3), 393–398.

Gladwell, M. (2019). *Talking to strangers: What we should know about the people we don't know*. New York, Little, Brown; Company.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, *108*(2), 419–453.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior and Organization*, *93*, 293–300.

Goldstone, R. L., & Chin, C. (1993). Dishonesty in self-report of copies made: Moral relativity and the copy machine. *Basic and Applied Social Psychology*, *14*(1), 19–32.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Granhag, P. A., & Strömwall, L. A. (2002). Repeated interrogations: Verbal and non-verbal cues to deception. *Applied Cognitive Psychology*, *16*(3), 243–257.

Grice, H. P. (1975). Logic and conversation (P. Cole & J. L. Morgan, Eds.). In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics vol. 3: Speech acts*. New York, Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA, Harvard University Press.

Hancock, J. T., & Toma, C. L. (2009). Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, *59*(2), 367–386.

Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How distance between truth and lies drives dishonesty. *Journal of Experimental Social Psychology*, *49*, 263–266.

Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *The American Economic Review*, *88*(4), 947–969.

Hurkens, S., & Kartik, N. (2009). Would i lie to you? on social preferences and lyinig aversion. *Experimental Economics*, *12*(2), 180–192.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kraut, R. E. (1978). Verbal and nonverbal cues in the perception of lying. *Journal of Personality and Social Psychology*, *36*(4), 380–391.

Kristal, A. A., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, *117*(13), 7103–7107.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807.

Lee, K. (2013). Little liars: Development of verbal deception in children. *Child Development Perspectives*, *7*(2), 91–96.

Levine, T. R. (2010). A few transparent liars: Explaining 54% accuracy in deception detection experiments. *Annals of the International Communication Association*, *34*(1), 41–61.

Levine, T. R. (2014). Truth-default-theory (tdt): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 1–15.

Levine, T. R. (2019). *Duped: Truth-default theory and the social science of lying and deception*. Tuscaloosa, AL, University of Alabama Press.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, Wiley.

Madigan, R., & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics*, *42*(3), 240–249.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology*, *89*(1), 137–149.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Montague, R., Navarro, D. J., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. Cognitive Science Society.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology*, *41*(1), 167–194.

Oey, L. A., Schachner, A., & Vul, E. (2022, May 25). Data for "Designing and detecting lies by reasoning about other agents". Retrieved from https://osf.io/x6rhs/.

Oey, L. A., Schachner, A., & Vul, E. (2019a). Designing good deception: Recursive theory of mind in lying and lie detection (A. K. Goel, C. M. Seifert, & C. Freksa, Eds.). In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society*, Montreal, QB. Cognitive Science Society.

Oey, L. A., Schachner, A., & Vul, E. (2019b). Recursive theory-of-mind in the design of deception: A rational model of lying and lie detection, Talk presented at the 45th Annual Meeting of the Society for Philosophy & Psychology, San Diego.

Oey, L. A., & Vul, E. (2021). Lies are crafted to the audience (T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible, Eds.). In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, Vienna, Austria, Cognitive Science Society.

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, *31*(4), 259–264.

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás and Endress (2010). *Psychological Science*, *26*(9), 1353–1367.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1-2), 122–136.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *4*, 515–526.

Ransom, K., Voorspoels, W., Navarro, D. J., & Perfors, A. (2019). Where the truth lies: How sampling implications drive deception without lying. *PsyArXiv*.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead other. *Journal of Personality and Social Psychology*, *112*(3), 456–473.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA, Harvard University.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in america: Three studies of self-reported lies. *Human Communication Research*, *36*(1), 2–25.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justification). *Psychological Science*, *23*(10), 1264–1270.

Shalvi, S., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management*, *22*, 16–27.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Stahl, D. O. (1993). Evolution of smartn players. *Games and Economic Behavior*, *5*(4), 604–617.

Street, C. N. (2015). Alied: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, *4*(4), 335–343.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*(8), 1023–1036.

Tyler, J. M., Feldman, R. S., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Social Psychology*, *42*, 69–77.

van't Veer, A. E., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making*, *9*(3), 199–206.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? meta-analyzing the effort of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition*, *7*, 462–469.

Vrij, A. (2008). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior*, *35*(10), 1323–1336.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142.

Vrij, A., Hartwig, M., & Granhag, P. A. (2019). Reading lies: Nonverbal communication and deception. *Annual Review of Psychology*, *70*, 295–317.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Withall, A., & Sagi, E. (2021). The impact of readability on trust in information (T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible, Eds.). In T. Fitch, C. Lamm, H. Leder, &

K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual meeting of the cognitive science society*, Vienna, Austria. Cognitive Science Society.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, *14*, 1–59.